Predicting What Will Happen When We Act. What Counts for Warrant?

Nancy Cartwright

Department of Philosophy, Logic, and Scientific Method, London School of Economics, Houghton Street, London WC2A 2AE, UK
and
Department of Philosophy, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92103, USA.

Abstract

To what extent do the results of randomized controlled trials inform our predictions about the effectiveness of potential policy interventions? This crucial question is often overlooked in discussions about evidence-based policy. The view I defend is that the arguments which lead from the claim that a program works somewhere to a prediction about the effectiveness of this program as it will be implemented here rests on many premises, most of which cannot be justified by the results of RCTs. RCTs only provide indirect evidence for effectiveness, and we need much more than just RCTs results to make reliable predictions.

Keywords: effectiveness, randomized controlled trial, warrant, evidence-based policy, argument

1. Introduction

My topic is *effectiveness hypotheses*: hypotheses that a well-described policy/treatment will work for us: it will result in an improvement in a well-specified outcome in a targeted situation implemented as we would implement it. Evidence-based policy (EBP) advocates have invested a great deal of effort in providing warehouses that evaluate and store evidence for effectiveness, warehouses to be visited by 'ordinary' policy makers and analysts, like the

Email address: N.L.Cartwright@lse.ac.uk (Nancy Cartwright)

Cochrane Collaboration for medical studies, the Campbell Collaboration for social policy, the US Department of Education's What Works Clearinghouse, or the Greater London Authority's Project Oracle for 'Understanding and sharing what really works' against youth violence.

These warehouses advertise that they store programs that 'work'. What they store are programs for which there is good reason to think they work somewhere; if we are lucky, in a few somewheres. The warehouse keepers police scientific studies that aim to establish causal connections between a program and a targeted outcome; their purchasing rules favor randomized controlled trial (RCT) study designs. A study like that provides *direct* evidence that the program worked there, then, in the study population. What makes that evidence that it will work here, now, as we would implement it?

Knowledge claims, including effectiveness predictions, are warranted by good arguments, arguments that are both valid and sound. *Warranted*: we have good reason to accept these claims; *valid*: the conclusion follows from the premises; *sound*: there is good reason to think the premises are true. The reminder that conclusions are justified by good arguments underlines two important facts usually underplayed in EBP discussions:

1. Evidence is a 3-place relation: e is evidence for h *relative* to an argument A for h. Failing the other premises in A, or relative to a different argument A′, e can be totally irrelevant to h.
2. Arguments are like chains: they are only as strong as their weakest premise. Focusing on the argument forces the premises to the fore. Often the unstated premises are the most dicey.

2. Warranting 'It will work for us'

So:

A well-established empirical claim e is evidence for hypothesis h relative to a good argument A (or A′, A″, A‴, . . .) only if e is a premise in A, which is itself a good argument for h (or, is a premise in A′ which is itself a good argument for a premise in a good argument A for h, etc.).

The EBP literature rates positive outcomes in well-conducted RCTs as gold standard evidence for effectiveness predictions. What's the argument?

One could read this as a question about 'external validity': Will the experimental result hold elsewhere, where there are conventionally cited facts, like similarity between target and experimental situations, that are supposed

to make this likely. But this is the wrong way to look at the relation between experimental results and the claims whose truth they bear on. Experimental results can help justify confidence that the same result – or that some different result – will hold elsewhere; i.e., they can be evidence for one of these claims. Whether they are evidence depends on whether they play the right kind of role in a good argument for that claim. Similarity, or just the right kind of dissimilarity, might play a role. But if so, only by fitting into a good argument.

What then might a good argument look like that makes RCT results evidence for effectiveness predictions?

RCT results are normally *effect sizes*: ES = the difference in the average of the outcome (y) in treatment group and in control group (av(y)_T - av(y)_C). Suppose these effects are not accidental but generated in accord with a causal principle obtaining in the study setting, say of this form:

$$\text{CP: y(i)} = \text{a} + \text{b(i)x(i)} + \text{z(i)}^1$$

where y(i) is the outcome for individual i in the study population, x(i) is the treatment variable, a is constant and z(i) represents all other causal clusters that contribute linearly with x to produce y in i. It is apparent from this principle that x is a genuine contributor to y for at least some i in this setting if and only if $\text{b(i)} \neq 0$. A well-known argument – labeled here *Argument A′*[2] – shows that, under usual assumptions characterizing ideal RCTs,

$$\text{ES} = \text{av(b)(X - X')}$$

where X is the value of the treatment in the treatment group and X', the value in the control group. If the effect size is non-zero, av(b) is non-zero, so b(i) is non-zero for some i[3] and thus x is a genuine contributor to y for some individuals in populations subject to CP.

So we have a good argument – A′ – that has among its premises:

- e: 'The effect size of x for y in a well-conducted RCT on populations subject to CP is non-zero.'

[1]Note: this is not a regression equation. It represents genuine causal relations, which regression equations will generally not do. I choose this simple linear form for convenience. The same conclusions hold for more complex nonlinear principles.

[2]Cf. (Cartwright, 2007, 15-16).

[3]But not conversely. Positive and negative individual effects can cancel to yield ES=0.

and as its conclusion,

- h_1: 'x contributes to the production of y for some individuals in populations subject to CP.'

Thus e is evidence for h_1 *relative* to argument A′ and thereby relative to the other premises in that argument (including the assumption that conducting the experiment well – randomizing, blinding, etc. – delivered the features required in ideal RCTs). To establish e's evidential relevance to h, we now need a good argument – A – where h_1 figures essentially as a premise and h as conclusion.

But note. Often CP is written with the i's implicit:

$$\text{CP': } y = a + bx + z$$

This suggests that b is constant. But there are few treatments for which this is likely. The treatment is usually only the salient factor in a cluster, and it produces a contribution *when all the factors in the cluster take the right values at once*; x contributes to y – but only when cooperating with other factors (sometimes called 'causal cofactors' or 'complementary component causes') and often a great many of these. In CP, b(i) represents the values of all these supporting factors in one fell swoop.

Now to the argument. First we need a properly formulated conclusion. Maybe...

- h_{ES}: 'If x = X in our setting, as opposed to x = X', keeping fixed all other causes of y here[4], the effect size would be ES here too'.

Will x make the same average contribution in the situation here as in the study situation there? Recall: b is not constant; and the effect size is its average. The average in each situation depends on the distribution of the supporting factors for x in that situation. Even if the same principles govern the two, that provides no reason for the distributions of support factors to be the same. To the contrary, this distribution often depends heavily on local circumstances.

Nor is the same distribution what you want. You'd really like to arrange to a distribution favouring values of b that provide the largest contribution.

[4]Excepting those downstream from x.

At the least, you want some values which make x's contribution positive and these should outweigh those making it negative. If getting negative contributions in some individuals is bad, then you want none of these.

Suppose though we aim to predict simply h_{cont}, that the policy will contribute positively in our situation. What can make RCT evidence relevant? Let 'x plays a causal role' mean x appears in the principle governing y. Here is what seems to be the weakest valid argument using RCT results there as a premise, concluding that the program contributes here.

> Argument ***A***:
>
> A1. x plays a causal role in the principle that governs y's production there.
>
> A2. x plays a causal role here as well as there.
>
> A3. The support factors necessary for x to operate are present for some individuals here.
>
> Therefore, x plays a causal role here and the support factors necessary for it to operate are present for some individuals.

The RCT enters in a different argument, supporting premise A1. It is not *direct* evidence for effectiveness, where a well-warranted empirical claim e is *direct evidence* for h if and only if e figures essentially in a good argument for h. A′ is a valid argument taking as premise a positive effect size in an experiment and as conclusion, that the program contributes to the outcome in the study situation. The other premises in A′ are about further features of the study (like 'confounding factors are distributed evenly in treatment and control groups'). Evidence warehouses police these premises for particular studies. Find a program in a conscientious warehouse provides good reason to think there is a valid and sound argument that x plays a causal role somewhere – which is the first premise in argument A.

So the RCT result can be evidence for effectiveness here, but *indirectly*. It is not a premise in an argument for effectiveness but a premise in an argument for a premise:

[Figure 1 goes here with the following caption: RCTs as indirect evidence for effectiveness.]

Its relevance is highly conditional, depending on the validity and soundness of arguments A′ and A. As in figure 2, a positive effect size in an RCT is leveraged into evidence that the program works there in the RCT by A′;

'it works there' is leveraged into evidence for 'it works here' by A. If either A or A′ fail, the lever drops and the evidential relevance disappears with a thud:

[Figure 2 goes here with the following caption: RCTs as conditional evidence for effectiveness.]

A′ and A are valid, so what matters is their soundness. We may grant that A′ is good if the program appears in a reputable warehouse. A2. and A3. are the additional premises necessary in A. What arguments support these? That's the problem. There are no warehouses for information like this and the kind of information needed is hard to come by. I don't see how A2. can be supported without a great deal of theory; so too with A3., in order to identify what the support factors *are*. A3. also requires local knowledge to determine if even some of the right values for support factors obtain let alone a desirable distribution.

Now consider two objections to my account of what counts as warrant for effectiveness predictions.

First: RCTs are often advocated by people who think our claims to theoretical knowledge are too slippery to trust. So they oppose my view about how A2. gets warranted. They offer an alternative: more RCTs, with as much variation in circumstances as possible. I agree. More RCTs, especially across a variety of circumstances, can improve the warrant for effectiveness predictions – because they support premises like A2.: the program plays a causal role here. How?

That's the rub. The argument could be by enumerative induction: swan 1 is white, swan 2 is white, ...; x plays a causal role in situation 1, x plays a causal role in situation 2, ... How good is that argument? Induction demands a large and varied inductive base – lots of swans from lots of places; lots of RCTs from different populations. It also requires that the observations be projectable, plus an account of the range across which they project. Electron charge is projectable everywhere – one good experiment can generalise to all; bird colour sometimes is; causality is dicey. Many causal connections depend on intimate, complex interactions among factors present so that no special role for the factor of interest can be prised out and projected to new situations.

Rather than some weak inductive argument, I urge a rigorous deductive argument. Then we know what we are betting on when we bet on the conclusion. So I add the premise that x's causal role is the same here as in all those other places. That makes the challenge clear: What is the warrant for

this very strong claim? This matters because of the weakest link principle: the conclusion can have no more warrant than any of its premises.

Second. Surely the best evidence that the program will work here is an RCT here. Agreed, this could at least be good evidence. *Could be*, were it possible. We never do RCTs on the same population at the same time. And both matter. A sample is almost never representative, that is: governed by the same causal principles and having the same probability distribution over causally relevant factors. And time cannot be ignored. Are the causes the same now as when the study was done? That's a serious question for socioeconomic policy since economists from J.S. Mill[5] to the econometrician David Hendry[6] worry that past regularities are poor guides to the future because the background arrangement of causes shifts so often and so unpredictably. Of course the experimental population could be representative enough and the causes stable enough. Let's just get this stated explicitly as one of our premises so that the need for warrant for it is transparent.

3. Conclusion

It is not a new idea that evidence is relative to an argument, and it may not be controversial. But taking it seriously matters. It is altogether too easy, when we do not keep the arguments to the fore, to overestimate the warrant our studies can deliver. RCTs for instance. Evidence-based policy takes them as gold standard evidence for effectiveness claims, though with a caution. The U.S. Department of Education, for example, warns that trials on white suburban populations do not constitute strong evidence for large inner city schools serving primarily minority students.[7]

This kind of warning conceals what needs to be exposed. What argument makes a particular RCT result evidence for a particular effectiveness prediction? If evidence, it is *indirect* – there are layers of arguments to get from study results to effectiveness conclusions:

[Figure 3 goes here with the following caption: Layers of arguments between RCT results and effectiveness predictions.]

They all have additional premises, every one of which is essential for the security of the conclusion. No matter how firm the RCT result, the

[5]Cf. (Mill, 1836 [1967]).

[6]Cf. (Hendry, 2004, 12-13) and references therein.

[7]Cf. (U.S. Dept. of Ed., 2003, 10).

effectiveness conclusion can have no greater claim to knowledge than the shakiest of these premises.

[Figure 4 goes here with the following caption: The chain from RCTs to effectiveness predictions.]

Nor is this unusual. Most of our knowledge claims, even in our securest branches of science, rest on far more premises than we like to imagine, and far shakier. This recommends a dramatic degree of epistemic modesty.

References

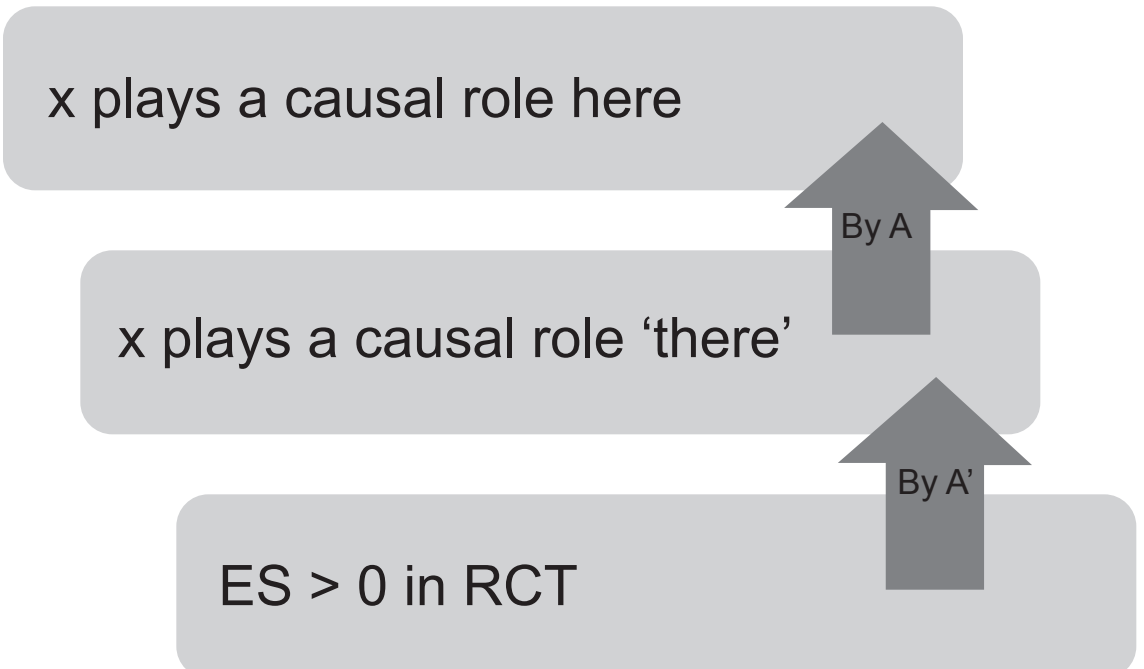
Cartwright N. Are RCTs the gold standard? BioSocieties 2007;2(1):11–20.

Hendry D. Causality and exogeneity in non-stationary economic time series. Causality: metaphysics and methods, CTR 18/04. London: CPNSS, 2004.

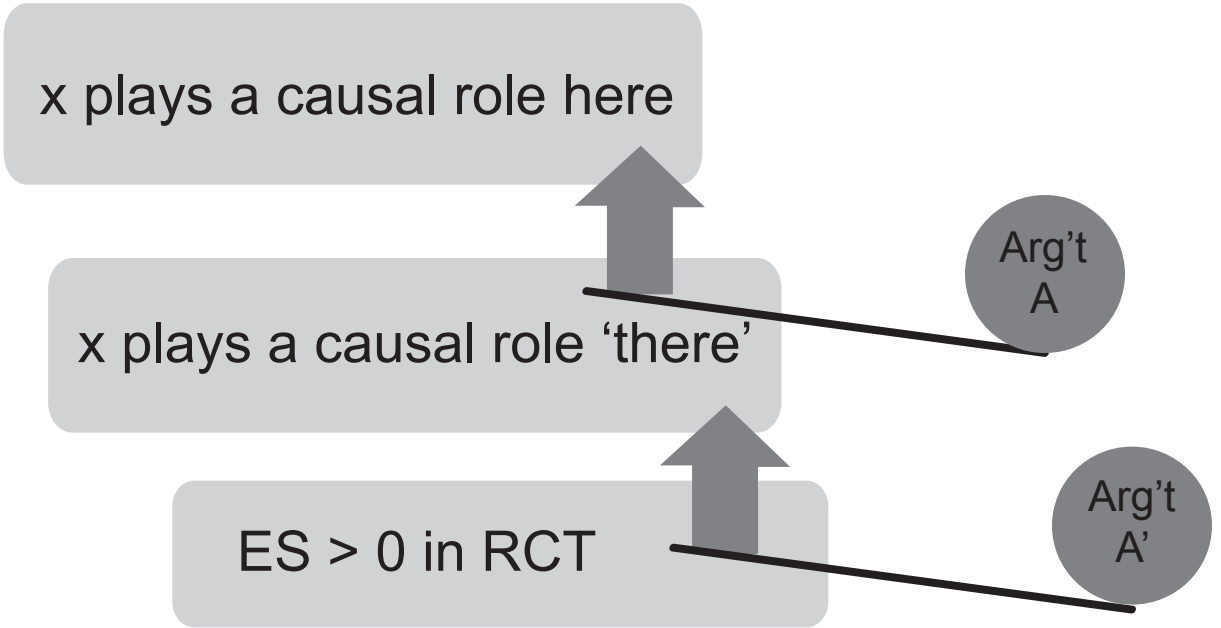
Mill JS. On the definition of political economy and on the method of philosophical investigation in that science. In: Collected Works of John Stuart Mill. Toronto: University of Toronto Press; volume 4; 1836 [1967]. .

U.S. Dept. of Ed. . Identifying and Implementing Education Practices Supported by Rigorous Evidence: A User Friendly Guide. Washington, D.C.: U.S. Department of Education, 2003.

The RCT as *indirect evidence* for effectiveness



The RCT as *conditional evidence* for effectiveness



It will produce improvements here

The helping factors for it are w, y, z

We have w, y, z here

It plays a causal role here

?

?

?

?

?

?

It plays the same role here and there

It plays a causal role here (and there)

?

?

R C T

R C T

It will produce improvements here

It plays a causal role here

It plays a causal role here (and there)